Automated Image Processing and Analysis Tools for DNA Origami Nanostructure Characterization by Atomic Force Microscopy

M. Chiriboga*, S. A. Díaz*, I. L. Medintz*

* Center for Bio/Molecular Science and Engineering, U.S. Naval Research Laboratory, 4555 Overlook Ave. S.W., Washington, DC 20375, USA

Abstract

The DNA origami technique is highly utilized in the DNA nanotechnolgy field for its ability to self-assemble geometric architectures with near atomic precision. The structural fidelity will vary depending on the assembly reaction conditions. Therefore, researchers have relied heavily on atomic force microscopy (AFM) for structural characterization. In particular for its ease of use, high throughput capacity, and ability to resolve distances on the sub-nanometer scale. As a result, enormous data sets of DNA origami images are being generated faster and larger than can be analyzed within acceptable error. Computational approaches must be employed to parse and extract meaning from the data. Here, we discuss some common open-source software used to automate the processing and analysis of DNA structures in AFM images.

Keywords: Atomic Force Microscopy, DNA Origami, DNA Nanostructures, Image Analysis, Machine Learning

Introduction

The field of DNA nanotechnology leverages the intrinsic biochemical and physical properties of DNA to form complex 2- and 3-dimensional nanoscale architectures [1]. Self-assembly techniques, such as DNA origami or DNA bricks, enable nearly any arbitrary nanoscale geometry to be assembled (Figure 1) [2–6]. Advantageously, DNA origami assemble on the molecular scale and thus the overall number of structures is proportional to Avogadro's number. Furthermore, the chemical properties of DNA allow for easy conjugation to a variety of functional materials. The sequence specificity of Watson-Crick base pair rules means functional group interactions can be predictably modulated through changes in the highly programmable DNA sequence [7]. Hence in practice, DNA origami can reliably arrange functional groups into logical arrays with spatial precision on the order of < 1 nm [8]. These properties make DNA origami a promising platform for the rapid prototyping of novel DNA based nanophotonic devices [9], dynamic DNA robots [10], and drug delivery vehicles [11, 12]. Although DNA origami is poised to underpin breakthroughs in the areas of healthcare and engineering, certain aspects of the technique still lack maturity and could benefit from community driven investment. DNA's sensitivity to environmental conditions (temperature, pH, ionic strength) means assembly optimization needs to occur in order to achieve suitable structure fidelity. Leading to somewhat *ad hoc* assembly protocols lacking the rigorous standardization necessary for objective inter-study comparison. To overcome, this researchers often report formation efficiencies from polyacrylamide or agarose gel electrophoretic separation based on molecular weight [13]. However gel separation is variable based on interpersonal technique, interpretation, and environmental conditions begetting a lack of rigor. Also gel separation is not analytic as two DNA origami may have distinct geometries but identical electrophoretic mobility. In response, researchers have turned to atomic force microscopy (AFM) and electron microscopy (EM) as primary structural characterization tools [14, 15]. In particular AFM has emerged as a preferential high-resolution imaging modality for bio-molecules due to its ease of use and relatively high throughput capacity.

AFM as an imaging metrology was first proposed by Binnig *et. al.* in 1986 [16]. Since then, AFM has become an essential to the field of DNA nanotechnology [15]. As a scanning probe microscopy (SPM) technique, AFMs reconstruct a 3D topographic mapping of a sample by raster-scanning a sharp nanometer sized probe across a sample's surface. The probe's tip is located at one end of a cantilever which is deflected by induced interactions between the tip and the sample surface. This bends the cantilever which can be measured through positional changes in the angle of a laser beam irradiating the opposite end of the cantilever [17]. This approach, called the optical lever method and is highly utilized for its favorable signal to noise ratio [15]. Thus, the method has become a standard for the super-resolution imaging of soft materials, such as DNA origami. To exemplify the power of this approach, Mou *et. al.* was able to resolve the periodicity of a B-form DNA helix to be 3.4 ± 0.4 nm, well below the Rayleigh-Abbe diffraction limit of visible light [18, 19]. Of particular note, the AFM can routinely achieve this resolution over the entire surface of a structure with no need for staining, site specific chemi-

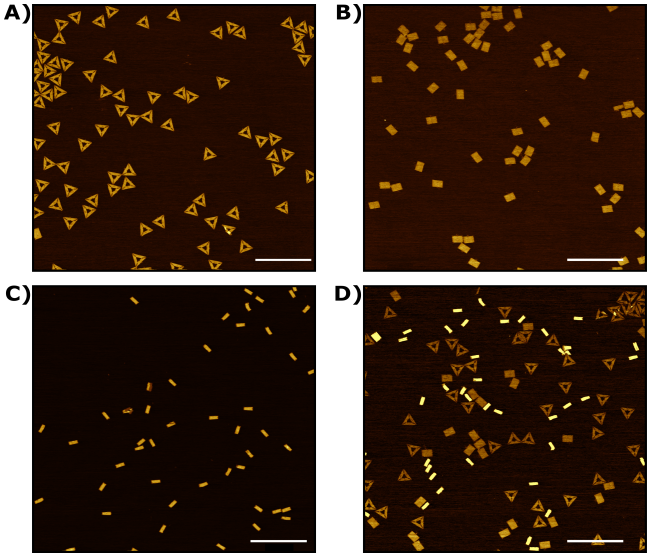

Figure 1: Selected images of DNA origami A) triangle, B) breadboard, and C) nanotube structures. Once fully assembled, structures can be combined to make D) heterogeneous populations. White scale bar (lower right) is 500 nm.

cal modification, nor ensemble averaging. Furthermore, there is no need for an electronically transmissible sample, ideal for biologicals. Each of these advantages exemplify constraints imposed on other imaging techniques such as EM and super-resolution fluorescence-based microscopy [20–22].

One commonality of all these microscopies is they enable acquisition of large image sets for statistical analysis. Each image potentially contains thousands of structures; for example, a 20×20 μm^2 AFM image of DNA origami on a mica substrate can easily contain over 10,000 origami assemblies. To garner statistics or quantitative metrics on structural fidelity in various conditions, multiple images from multiple experiments must be examined. This tedious labor of identifying, classifying, and analyzing objects in static images is a near ideal case for computer automation. But, compared to other microscopy communities, the bio-AFM community has scant open-source and freely available software empowering the automated identification DNA nanostructures. At present, AFM analysis requires initializing analysis pipelines *de novo* or relying on manual annotation [30]. The former is not cost effective and may not always be feasible for first pass analyses of diverse samples sets. While the latter is typically inconsistent between annotators and even between multiple annotations performed by a single annotator. Additionally, both require substantial time investment and often are not directly extensible, limiting objective inter-study comparison and reproducibility. One could argue this landscape is ripe for innovation as machine learning and neural network tools grow more ubiquitous. To offer perspective, we provide some review of two open source software tools commonly used for AFM image analysis and processing (Table 1).

Available Software

ImageJ

One of the oldest image analysis tools, widely adopted across the biological sciences, is ImageJ (Table 1) [23]. Written in the Java language, ImageJ was created to provide a simple and intuitive user interface for researchers lacking experience in computer science and/or coding. A major advantage of ImageJ is its augmentation through plugins, *i.e.* separately installed modular software components created to enhance or add specific functionality. This has separately attracted computer programmers who identified plugins as an opportunity to develop new software. Coupled with the open-source nature of ImageJ, a vibrant ecosystem of community driven functionality has emerged which continues to exist to this day [24]. For example, ImageJ was used to identify and assay drug dependent microtubule stabilization in AFM images [31]. In another case, Boudaoud *et. al.* developed a plugin (FibrilTool) enabling the analysis of fibrillary structures in raw AFM images [32]. Similarly, Bangalor *et. al.* used ImageJ to measure bending in a Glycosylase-DNA complex through AFM images to develop a DNA lesion sensing method [33]. In 2011 the Fiji distribution of ImageJ was released and was specifically targeted to biological image analysis [24]. The distribution was bundled with a variety of plugins attempting to streamline the installation and configuration of, at times, complicated software. Furthermore, developers implemented version control, issue tracking, and support for a broader range of scripting languages (Jython, JRuby, Beanshell) [24]. For these reasons Fiji has since become one of the most popular tools for biological image analysis [34].

Gwyddion

Gwyddion is an open-source modular software platform focusing on the processing and analysis of SPM data including AFM (Table 1) [25]. The codebase is written in the C language and separated into libraries and modules. The libraries, provide the binary for basic interfaces, functionalities, and data structures while the modules contain most of the analysis tools including data processing methods, interactive tools, and graph methods [25]. Gwyddion is designed to handle diverse input data, delivering a consistent user experience for a variety of SPM instruments. However, the software can specifically model AFM imaging process effects such as tip convolutions, feedback loops, noise smoothing, and simulation of select artifacts or parameters which can affect cantilever stiffness [35]. For example, Gwyddion

Table 1: Selected AFM Image Processing and Analysis Software.

Software	Source Status	Price (USD)	Developer Affiliations	Ref
ImageJ	Open	Free	NIH	[23, 24]
Gwyddion	Open	Free	Czech Metrology Institute	[25]
WSxM	Closed	Free	NanoTec	[26]
Nanoscope Analysis	Closed	Free (Legacy)	Bruker	[27]
MountainSPIP	Closed	Varies	Digital Surf	[28]
FemtoScan Online	Closed	400-1,300	Advanced Technologies Center	[29]

has recently been used to characterize hydrogels [36–38] and DNA nanostructures [39–41] with nanometer resolution.

In ImageJ and Gwyddion, simple image processing algorithms can quickly be applied to a set of images through stacking. But, the real power is derived from the ability to automate analysis pipelines through complex scripting [30]. However, these procedures require discreet input variables or boundary conditions manually defined by the user. Moreover, scripts require optimization based on image quality, magnification, sample type, and experimental conditions. Recently, more developed toolkits have sought to overcome the limits of these software. TopoStats, a program for the automated tracing of bio-molecules in AFM images, aims to identify a range of molecules in without user input [30]. Though implemented Python, the TopoState codebase is built on top of Gwyddion and bundles the functionality of Gwyddion's methods with the versatility of common pythonic libraries (NumPy [42] and SciPy [43]). This permits a near plug and play data analysis pipeline [30]. TopoStats illustrates the demand for simple yet advanced image analysis software. Therefore, with the increasingly ubiquitous nature of machine learning algorithms, the next breakthrough technology will likely be an intelligent one.

Convolutional Neural Networks

Modern deep convolutional neural networks (CNNs) been demonstrated to excel at object identification tasks [44]. Additionally, advancements in both software and hardware have allowed for CNNs to carry out these tasks rapidly and *en masse* [45]. For example in the 1980s, early CNNs were originally employed to automate recognition of hand written zip codes for the US postal service [46, 47]. More powerful modern CNNs have been used to analyze radiological (X-ray, CT, MRI, etc.), histological, ultrasound, and endoscopic images with high accuracy [48]. Hence, DNA origami identification is a well suited task for CNNs. One example of CNN analysis of AFM images is by Bai and Wu,[49] who utilized the You Only Look Once (YOLO) CNN framework to identify DNA nanowires of different morphology. They demonstrated automated detection of DNA nanowires with varied morphology and observed 90% reliability in their test set at nanometer resolution [49, 50]. Similarly, Sotres *et. al.* combined YOLO and Siamese networks to identify and track DNA structures in real time using automated AFM imaging [51]. While these are two predominant examples of using CNNs to identify DNA structures via AFM, there are other examples utilizing deep learning in tandem with AFM to analyze non-DNA materials (organics [52], nanoparticles [53], proteins [54]). However, these studies may rely on force spectroscopy rather than topographic imaging meaning the tools are not always directly translational [52]. With the increased availability of open-source, high-performance, CNNs there is an opportunity for adoption in the DNA-AFM community where automated analysis tools are in high demand and supply is not yet being met.

Future Outlook

As the DNA origami technique continues to mature, the amount of annually collected AFM image data will likely increase in tandem. Furthermore, increasingly complex structures will require more intensive characterization. Before long, the comprehensive analysis of DNA origami in AFM images will be functionally impossible without rigorous automated analysis tools. Luckily, intra-image object identification is a perfect task for computational automation. Current freely-available software rely mainly on user-defined scripting perhaps suitable for industrial pipelines where operating procedures are standardized. But, in basic research, scripts require constant adjustment to accommodate experimental modification. However, machine learning and CNNs provide an opportunity to develop DNA specific tools to overcome these limits. For example, CNNs could be trained on standardized datasets, used to benchmark assembly protocols, and network weights can be distributed throughout the field. With wider adoption, a fully developed CNN or set of set of CNNs could serve to eliminate operator bias in AFM analysis, reduce time and resource waste from current inefficient methods, and most importantly enable direct and objective

inter-study comparison.

References

1. Seeman, N. C. *Journal of theoretical biology* **99** (1982).
2. Benson, E. *et al. Nature* **523** (2015).
3. Douglas, S. M. *et al. Nature* **459** (2009).
4. Douglas, S. M. *et al. Nucleic Acids Research* **37** (2009).
5. Rothemund, P. W. *Nature* **440** (2006).
6. Veneziano, R. *et al. Science* **352** (2016).
7. Watson, J. D. & Crick, F. H. *Nature* **171** (1953).
8. Huang, J. *et al. Small Structures* **1** (2020).
9. Bui, H. *et al. Advanced Optical Materials* **7** (2019).
10. Simmel, F. C. *Current Opinion in Biotechnology* **23** (2012).
11. Linko, V., Ora, A. & Kostiainen, M. A. *Trends in Biotechnology* **33** (2015).
12. Weiden, J. & Bastings, M. M. *Current Opinion in Colloid & Interface Science* (2020).
13. Mathur, D. & Medintz, I. L. *Analytical Chemistry* (2017).
14. Jungmann, R., Scheible, M. & Simmel, F. C. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **4** (2012).
15. Main, K. H. *et al. APL Bioengineering* **5** (2021).
16. Binnig, G., Quate, C. F. & Gerber, C. *Physical Review Letters* **56** (1986).
17. Alexander, S. *et al. Journal of Applied Physics* **65** (1989).
18. Born, M. & Wolf, E. (2013).
19. Mou, J., Czajkowsky, D. M., Zhang, Y. & Shao, Z. *FEBS Letters* **371** (1995).
20. Green, C. M. *et al. Nanoscale* **9** (2017).
21. Jungmann, R. *et al. Nano Letters* **10** (2010).
22. Mathur, D. *et al. Nanoscale* **11** (2019).
23. Rasband, W. S. Web Page. 2011. `http://imagej.nih.gov/ij/`.
24. Schindelin, J., Rueden, C. T., Hiner, M. C. & Eliceiri, K. W. *Molecular Reproduction and Development* **82** (2015).
25. Nečas, D. & Klapetek, P. *Open Physics* **10** (2012).
26. Horcas, I. *et al. Review of Scientific Instruments* **78** (2007).
27. Digital Surf. Web Page. `digitalsurf.com/software-solutions/scanning-probe-microscopy/`.
28. Bruker Instruments. Web Page. `nanoscaleworld.bruker-axs.com/nanoscaleworld/forums/t/812.aspx`.
29. Advanced Technologies Center. Web Page. `nanoscopy.net/catalog`.
30. Beton, J. G. *et al. Methods* (2021).
31. Chiang, Y.-L. *et al. 42nd Society for Biomaterials Annual Meeting and Exposition 2019* (2019).
32. Boudaoud, A. *et al. Nature Protocols* **9** (2014).
33. Bangalore, D. M. *et al. Scientific Reports* **10** (2020).
34. Walter, T. *et al. Nature Methods* **7** (2010).
35. Klapetek, P. & Nečas, D. *Measurement Science and Technology* **25** (2014).
36. Johns, M. A. *et al. ACS Omega* **3** (2018).
37. Paul, A. *et al. Soft Matter* **13** (2017).
38. Ţălu, Ş. *Polymer Engineering & Science* **53** (2013).
39. Bose, K., Lech, C. J., Heddi, B. & Phan, A. T. *Nature Communications* **9** (2018).
40. Pyne, A., Thompson, R., Leung, C., Roy, D. & Hoogenboom, B. W. *Small* **10** (2014).
41. Shu-wen, W. C., Banneville, A.-S., Teulon, J.-M., Timmins, J. & Pellequer, J.-L. *Nanoscale* **12** (2020).
42. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. *Computing in Science & Engineering* **13** (2011).
43. Virtanen, P. *et al. Nature Methods* **17** (2020).
44. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521** (2015).
45. Steinkraus, D., Buck, I. & Simard, P. in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (2005).
46. Denker, J. S. *et al.* in *Advances in Neural Information Processing Systems* (1989).
47. LeCun, Y. *et al. Neural Computation* **1** (1989).
48. Lee, J.-G. *et al. Korean Journal of Radiology* **18** (2017).
49. Bai, H. & Wu, S. *Nanotechnology and Precision Engineering* **4** (2021).
50. Bai, H. & Wu, S. *Microscopy and Microanalysis* **27** (2021).
51. Sotres, J., Boyd, H. & Gonzalez-Martinez, J. F. *Nanoscale* (2021).
52. Alldritt, B. *et al. Science Advances* **6** (2020).
53. Mencattini, A. *et al.* in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (2018).
54. Boginskaya, I. *et al. Applied Sciences* **9** (2019).